Representation Learning for Cross-Embodiment Inverse Reinforcement Learning from Mixed-Quality Demonstrations

Anonymous authors
Paper under double-blind review

Abstract

We study the problem of cross-embodiment inverse reinforcement learning, where we wish to learn a reward function from video demonstrations in one or more embodiments and then transfer the learned reward to a different embodiment (e.g., different action space, dynamics, size, shape, etc.). Learning reward functions that transfer across embodiments is important in settings such as teaching a robot a policy via human video demonstrations or teaching a robot to imitate a policy from another robot with a different embodiment. However, prior work has only focused on cases where near-optimal demonstrations are available, which is often difficult to ensure. By contrast, we study, analyze and define the setting of cross embodiment reward learning from mixed-quality demonstrations. We demonstrate that prior work struggles to learn generalizable reward representations when learning from mixed-quality data. We then propose and analyze several techniques for representation learning that are designed to enable effective cross-embodiment learning. Our results also give insight into how different representation learning techniques lead to qualitatively different reward shaping behaviors.

1 INTRODUCTION

Inverse reinforcement learning (IRL) (Arora & Doshi, 2021) seeks learn a reward function from observed agent behavior. However, the field of imitation learning (Hussein et al., 2017) has developed numerous techniques for direct policy learning from observed agent behavior. So why learn a reward function? From the earliest days of IRL research, Ng et al. (2000) and others have argued that reward functions provide a succinct description of behavior. The idea being that if we can learn a reward function from observing an agent in one task, it should be the case that we can use that reward function to help teach the same task to agents with different embodiments (e.g., different action space, dynamics, size and shape, etc.). The idea of allowing agents with different embodiments to learn from each other is typically called *cross-embodiment* (Zakka et al., 2022) or *cross-domain* (Niu et al., 2024) learning. In this paper we focus on cross-embodiment IRL, with the goal of learning robust reward functions that can transfer across different embodiments.

Cross-embodiment IRL would enable robots to learn rewards from watching humans perform tasks and would allow robots and other AI agents to learn by watching other agents. However, given an embodiment mismatch between the demonstrator and the learner, we cannot simply imitate the actions of the demonstrator since the action spaces may be completely different. Furthermore, the actions are typically unavailable when learning only from video observations (Torabi et al., 2019). Learning an embodiment-independent reward function is a compelling solution to the problem of cross-embodiment policy learning as it should allow an agent of any embodiment to learn how to perform the desired task via reinforcement learning. However, prior work has shown that reward functions learned from demonstrations are often entangled with the dynamics, making cross embodiment transfer difficult (Fu et al., 2017).

Recently, Zakka et al. (2022) developed a novel approach for cross-embodiment reward learning from video demonstrations. Using near-optimal demonstrations across several different embodiments they first learn an embodiment-invariant representation using temporal cycle-consistency (Dwibedi et al., 2019). The base assumption this method uses is that there is a temporal similarity between the data it is trained on, i.e., there are similar frames or checkpoints that each video demonstration will share with another. The reward is then formulated as the distance between the current state embedding to that of a goal embedding and this learned reward is optimized via RL to achieve generalization to an unseen embodiment. However, one of the main limitations of this recent breakthrough by Zakka et al. (2022) is that, in order to ensure temporal similarity when performing representation learning, the approach requires *near-optimal demonstrations across each embodiment*. Prior work has shown that human demonstrations and other interactions with AI systems can be noisy (Chuck et al., 2017; Mandlekar et al., 2022), irrational (Chan et al., 2021; Ghosal et al., 2023), and sometimes adversarial (Wolf et al., 2017; Jagielski et al., 2018; Oravec, 2023). Thus, assuming that demonstration data is near-optimal is unlikely to be true in practice.

We propose an analyze several different approaches for perform cross-embodiment IRL from *mixed-quality demonstrations*. We study three different approaches: (1) *Vanilla RLHF* where we assume access to high-level preferences over demonstrations in our training dataset and use these preference labels to perform reinforcement learning from human feedback (Christiano et al., 2017); (2) *X-Prefs*: We extend prior work by Zakka et al. (2022) to enable learning an embedding function via trajectory preferences and then minimizing the distance between the learned embedding and a goal embedding; and (3) *TCC-Buckets*: A method that seeks to apply a temporal cycle-consistency representation learning to mixed-quality data by leverage high-level knowledge of the relative goodness of demonstrations by first binning the demonstrations into several "buckets" or groups based on ordinal labels denoting their goodness and then performing temporal cycle-consistency representation learning within each bucket.

The primary contributions of this work are: (1) We propose and formalize the new problem of cross-embodiment inverse RL from mixed-quality data; (2) We propose several new algorithmic approaches for this problem setting that build on and combine ideas from representation learning and preference learning; (3) We empirically study the RL performance, learned rewards and learned representations when learning with mixed data and show that prior approaches fail to perform well in this setting; (4) We provide empirical evidence that approaches that leverage information about the relative quality of the data are able to often learn transferable representations and corresponding rewards that transfer across embodiments even when learning from mixed-quality demonstrations.

2 PRIOR WORK

Imitation Learning from Observation Our work seeks to leverage video observations from one embodiment and learn to transfer these policies to new embodiments. Thus, our work falls within the general area of imitation learning from observation Torabi et al. (2019). Early work on inverse reinforcement learning learned reward functions based on state observations of demonstrations (Abbeel & Ng, 2004; Ziebart et al., 2008). Other work proposed imitation learning from state observations via model-based behavioral cloning (Torabi et al., 2018a) by learning an inverse dynamics model and subsequent work proposed to explicitly minimize inverse dynamics disagreement (Yang et al., 2019). However, these prior works focused on learning from low-dimensional state observation sequences. More recent work studies cases where the goal is to learn from video observations (Torabi et al., 2018b; Liu et al., 2018; Salimans & Chen, 2018; Goo & Niekum, 2019; Kidambi et al., 2021).

Cross-Embodiment Reward and Policy Learning Much prior work has focused on cross-embodiment learning (Niu et al., 2024) when given full access to the demonstrating agent's statespace. Fu et al. (2017) demonstrate that learned reward functions are often entangled with embodiment dynamics and propose an adversarial reward learning approach that seeks to learn an unshaped, state-only reward and Fickinger et al. (2021) perform cross-domain imitation learning based on optimal transport but both methods are restricted to low-dimensional state spaces rather than video observations. Recently, researchers have proposed methods that can scale to settings with partial observability, where only video observations of demonstrations are available. Zakka et al. (2022) propose a method for cross-embodiment inverse reinforcement learning from videos by finding correspondences across time across multiple videos. Xu et al. (2023) consider an alternative approach that leverages optimal human and robot demonstrations to learn resuable skills from videos. Other work seeks to train robot foundation models that can then be fine-tuned on arbitrary robot embodiments (Padalkar et al., 2023). Prior work on imitation learning from observations and cross-embodiment reward and policy learning focuses on learning from near-optimal demonstrations. By contrast, we seek to perform cross-embodiment learning from mixed-quality data.

Learning from Suboptimal Demonstrations: When ground-truth rewards are known, it is common to initialize a policy using demonstrations and then improve this policy using reinforcement learning (Hester et al., 2018; Gao et al., 2018; Wilcox et al., 2022). However, these methods typically do not consider embodiment mismatches and rely on designing a reward design which can easily lead to unintended behaviors (Ng et al., 1999; Amodei et al., 2016; Booth et al., 2023). Other work learns from demonstrations that are labeled good or bad (Grollman & Billard, 2011; Shiarlis et al., 2016) or are robust to a small number of unlabeled, poor demonstrations (Zheng et al., 2014; Choi et al., 2019). However, these prior works do not consider embodiment mismatch and require low-dimensional state observations. Prior work has considered learning reward functions from pairwise preferences over trajectories (Wirth et al., 2017; Christiano et al., 2017) and has shown the ability of these methods to extrapolate beyond the performance of suboptimal demonstrations (Brown et al., 2019; 2020). While there has been substantial progress recently in learning from pairwise preferences (Lee et al., 2021; Park et al., 2021; Bobu et al., 2023; Rafailov et al., 2024; Myers et al., 2022; Shin et al., 2023; Liu et al., 2023; Wilde et al., 2021; Ouyang et al., 2022), prior work does not consider the cross-embodiment setting that we explore in this paper.

3 PROBLEM FORMULATION

Our problem setting is inspired by Zakka et al. (2022), who consider cross-embodiment IRL. However, in contrast to Zakka et al. (2022), we seek to learn from mixed-quality, mixed-embodiment data. We investigate the problem of learning an agent-agnostic representation of a task, T, given a dataset of videos depicting agents performing the task. Formally, we define a dataset D as a collection of state-only video demonstrations $D = \{v_0, v_1, \ldots, v_i\}$, where each video contains a sequence of frames (2D images), $v_i = \{v_i^0, v_i^1, \cdots, v_i^j\}$ that depict the agent executing the task. In contrast to prior work (Zakka et al., 2022), we consider a set of *mixed-quality* demonstrations, where it is no longer guaranteed that agents will reach the goal state at the end of every demonstration. We refer to the demonstration set as *mixed-embodiment* if it contains demonstrations from more than one agent embodiment performing the same task.

Problem Statement: Given a task, T, and a mixed-quality, mixed-embodiment (MQME) demonstration dataset, D, can we extract a embodiment-agnostic approximation, $\hat{r}$, of the ground truth reward function using inverse reinforcement learning? Furthermore, can this learned reward function be used to successfully accomplish task T by performing RL on $\hat{r}$ with a new embodiment?

4 PRELIMINARIES

In this section we describe in detail the XIRL algorithm, an unsupervised method of learning embodiment-agnostic representations of tasks proposed by Zakka et al. (2022). XIRL assumes access to a video dataset, $D = \{v_0, v_1, \ldots, v_i\}$, of successful video demonstrations, v_i, for a task. Each video demonstration is an observation-only video (demonstrator actions are unobserved) containing a sequence of frames, $v_i = \{v_i^1, v_i^2, \cdots, v_i^j\}$. XIRL assumes that D contains video demonstrations from multiple agent embodiments, all performing the same task.

To learn from multi-embodiment data, Zakka et al. (2022) seek to learn a useful representation for cross-embodiment learning that aligns task progress in one embodiment to the corresponding task progress in a different embodiment. To address this, XIRL relies on temporal cycle-consistency (TCC) (Dwibedi et al., 2019) learning to establish task-aligned feature representations of cross embodiment demonstrations that captures information about the task itself, rather than the agent executing the task. The goal of TCC is to train an encoder, ϕ, that takes as input an image, s, and outputs an embedding vector $\phi(s)$. One of the primary benefits of XIRL's use of TCC is that it does not require embodiment labels and is completely unsupervised. First, random mini-batches of videos are sampled from D. Given ϕ, each video can be represented as a sequence of embedded images, $V_i = \{\phi(s_i^1), \phi(s_i^2), \cdots, \phi(s_i^{L_i})\}$, where $L_i = |v_i|$. Given a mini-batch of embedded videos, ϕ is updated by taking pairs of sequences V_i and V_j and computing a TCC Loss which aligns a random frame index, t, in V_j to the corresponding soft-nearest-neighbor frame, t', in V_i by minimizing the mean-squared error between the frame indices, $L_{ij}^t = (t' - t)^2$ (Zakka et al., 2022).

Following the self-supervised training of the encoder ϕ, XIRL grounds an embodiment-agnostic reward function to the demonstration set by computing a *goal embedding*. Because XIRL assumes that the demonstration set provided are always near-optimal, the last frame of every sequence, $v_i^{L_i}$, can be assumed to represent a state where the agent successfully completed the task. Therefore, the average of these final frames' embeddings are averaged to create a goal state embedding, $g = \frac{1}{N}\sum_{i=1}^{N} \phi(v_i^{L_i})$. Finally, both the encoder and the goal embedding are used to provide a reward signal during reinforcement learning. Specifically, the reward is the negative signed distance to goal, $r(s) = -\frac{1}{\kappa} \cdot \|\phi(s) - g\|_2^2$, where κ is a scaling parameter.

In the following sections, we build on the foundational work of Zakka et al. (2022) to study cross-embodiment learning when multi-embodiment demonstrations are of mixed quality and may not always successful complete the desired task.

5 METHODS

We seek to learn reward representations that generalize to unseen agent embodiments, In contrast to XIRL (Zakka et al., 2022), we assume that our dataset is MQME: *mixed-quality*, where the quality of the demonstration with respect to task success varies and also *mixed-embodiment* (i.e. demonstrations will be given by agents with various physical embodiments and action spaces). Dealing with mixed-quality data poses a challenge for two of the main components of the XIRL pipeline: (1) XIRL pretrains the video frame encoder ϕ with TCC which assumes that temporal alignment exists between all demonstrations, e.g., each demonstration starts in a similar start state configuration and is assumed to end at state that corresponds to task success, with several key corresponding intermediate steps. However, if some demonstrations are of mixed quality, this temporal alignment may not exist between pairwise samples of videos. (2) The reward formulation for XIRL depends on a reliable goal approximation, g, which is the average embedding from the final frame of all videos in the dataset. By removing the guarantee that all demonstrations successfully complete the task, the XIRL reward function may not guide the agent towards task completion as the goal embedding may no longer be reliable. In the following sections, we address these concerns and propose and discuss three approaches to performing cross-embodiment IRL from MQME data.

5.1 Reinforcement Learning From Human Feedback (RLHF)

The simplest way to address the issues of mixed-quality data in XIRL is to try learning a reward end-to-end with supervised Reinforcement Learning From Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), by having a human provide preference labels over an offline dataset of trajectories (Brown et al., 2019; Shin et al., 2023). In this approach, the demonstration dataset is augmented by a set of pairwise preference labels over the data. For a pair of demonstrations, (v_i, v_j), the notation $v_i \prec v_j$ indicates a preference of demonstration j over demonstration i. Using these labels, we follow prior work by employing a deep neural network, namely a reward predictor, $\hat{r}$, that maps video frames into a single real-valued reward that can be trained via backpropagation using the standard Bradley-Terry (Bradley & Terry, 1952) and Luce-Shepperd (Luce, 2005; Shepard, 1957) model Christiano et al. (2017) (see Appendix A for more details).

The benefit of using a vanilla RLHF approach is that it directly learns the reward end-to-end, with no intermediate latent representation or goal embedding used to manually compute the reward. Although this method requires some human burden compared to XIRL, using human supervision can lead to better and more human-aligned representations (Bobu et al., 2023; Tian et al., 2024) than purely unsupervised representation learning approaches such as XIRL.

5.2 XPrefs

Our second approach is motivated by existing inductive bias in XIRL. We introduce Cross Preference Learning (XPrefs). An approach to cross-embodiment learning that uses the same underlying architecture and reward function as XIRL, but with supervised learning from pairwise preferences. When provided with MQME data, we hypothesize that representation learning via TCC will fail to learn a correct embedding because both videos may not share the same task-relevant keyframes. Thus, we propose the use of preferences to learn a better latent embedding that can be used to guide RL via the same reward function used in XIRL, $r(s) = -\frac{1}{\kappa} \cdot \|\phi(s) - g\|_2^2$. To do this, we simply optimize $\phi(s)$ to maximize the likelihood of the preference labels, where

$$P(v_1 \prec v_2) = \frac{exp \sum_{s \in v_2} -\|\phi(s) - g\|_2^2}{exp \sum_{s \in v_1} -\|\phi(s) - g\|_2^2 + exp \sum_{s \in v_2} -\|\phi(s) - g\|_2^2} . \tag{1}$$

Note, the $\frac{1}{\kappa}$ term is removed from the reward function during reward learning, as the scaling term was added by XIRL to benefit the RL part of the pipeline. The equation above requires a known or calculated goal embedding, g. We assume that in addition to the MQME dataset, we have direct and privileged access to a known set of goal states, $G^* = \{g_1^*, g_2^*, \cdots, g_N^*\}$. Following XIRL, our goal embedding is simply the average embedding over all goals in G^*, resulting in $g = \frac{1}{|G^*|} \sum_{g_i^* \in G^*} \phi(g_i^*)$.

The introduction of this goal embedding into the softmax probability function (Eq. (1)) results in an "chicken-and-egg" cyclic dependency where both the embedding of the demonstration, $\phi(s)$, and g, which is a function of ϕ, will change every time the model updates. We theorize that fixing an arbitrary point in the embedding space as our goal state is appropriate and even beneficial for learning an embodiment-independent state representation as it is akin to fixing an origin for the space and fitting the state distribution around it in a way that is locally optimal. To fully study the efficacy of XPrefs, we explore both dynamic and static goal representations in Appenidx C and show that a static goal representation results in the best performance for reward learning.

5.3 Bucket-TCC

The third method we study, takes advantage of the unsupervised nature of TCC but also adapts the algorithm for a MQME dataset. We propose Bucket-TCC, which partitions the dataset into a number of "buckets" based on performance (See Appendix B for algorithm). We assume that a human labeller categorically assigns the trajectories into a bucket based on perceived performance. We then train a representation by applying a TCC loss only amongst trajectories within the same bucket and use this representation as a reward function following the same procedure as XIRL. Compared to XPrefs and RLHF, which require pairwise labels over a dataset, Bucket-TCC requires ordinal categorization which is far less burdensome in terms of the number of times human queries.

6 EXPERIMENTS

In this section, we seek to answer the following questions: (1) How does the quality of demonstrations degrade the learned reward and XIRL? (2) Can we learn good rewards and embeddings from MQME data? (3) How do the different methods described above perform when learning from MQME data?

6.1 Experimental Setup

Domain We conduct a series of experiments targeted at answering the aforementioned questions. For our experiments we use the X-MAGICAL imitation learning benchmark from Zakka et al. (2022). An example of this task is shown in Figure 2. The task involves pushing a set of blocks into the pink endzone using an agent of four possible embodiments: shortstick, mediumstick, longstick and gripper. The three stick embodiments all have the same action space but differ in their length, making the task easier for the longer embodiments and leading to qualitatively different optimal policies depending on the embodiment. Gripper not only has a different shape, but also has an extra action it can use to grip blocks with the pair of pincers on the agent.

MQME Data To simulate a mixed-quality, mixed-embodiment dataset for our experiments, we took policies trained via RL on the ground-truth reward (number of blocks pushed to the goal divided by 3) for each of the embodiments listed above and then degraded these pretrained oracle policies by iteratively adding randomness to the action selection. This type of noise injection is inspired by prior work that found it resulted in diverse suboptimal behaviors (Brown et al., 2020; Tien et al., 2022). There are a total of 200 trajectories for each embodiment (160 training and 40 validation trajectories for a training:validation split of 4:1). The dataset is evenly partitioned based on the number of blocks that are pushed in by the agent. By contrast, the X-MAGICAL dataset provided by Zakka et al. (2022) has the same embodiments as our MQME dataset but consists of approximately 1000 trajectories, almost 5 times our MQME dataset, per embodiment (split into training and testing sets), all of which are exclusively successful demonstrations of the task.

The number of preference labels used for training the reward model for RLHF and XPrefs is 5000, all of which were obtained by sampling them from a larger set of procedurally generated preferences by comparing all the pairs of trajectories in our MQME dataset. The preferences were generated according to which trajectory in a pair of trajectories from the dataset has the higher average ground truth environment reward per step over the length of the trajectory. The reason the average reward per step was used is due to the longstick embodiment having a shorter time horizon for the task as it is significantly easier to complete the task with that particular embodiment. The synthetic preferences were meant to loosely mimic how a human would provide preferences observing the task.

6.2 Baselines

In this section we describe the three baselines we compare against. All these baselines and methods use the same Soft Actor Critic code used in the original XIRL work: (1) **Reinforcement Learning on Ground Truth Reward.** As an oracle, we compare cross-embodiment IRL with running RL on the ground-truth reward from Zakka et al. (2022). At each step, the ground truth reward describes the fraction of total available blocks that are currently in the goal zone. (2) **XIRL Trained on X-MAGICAL.** This method uses the full pipeline of XIRL as described in Section 4 trained on the dataset of 200 successful demonstrations for each embodiment. This acts as an oracle since it provides the current-state-of-the-art performance for cross-embodiment IRL, but assumes access to near-optimal demonstrations for each embodiment. (3) **XIRL Trained on Mixed Data.** This method follows the XIRL pipeline but uses MQME data for TCC representation learning. The second step of the XIRL pipeline, goal embedding computation, is done with the same set of positive goal state examples we assume we have access to for the original methods developed in this paper. Thus, this baseline allows us to test the effect of mixed-quality data on XIRL and provides the main, non-oracle, baseline which we hope to significantly outperform.

6.3 Cross-Embodiment Learning from Mixed-Quality, Mixed-Embodiment Data

To study how well our proposed approaches and baselines compare when evaluated on MQME data, all the policies (except the oracle ground-truth RL baseline) were trained on 3 out of the 4 embodiments. We then evaluated the down-stream RL performance on the medium-stick held-out embodiment. To evaluate generalization performance, we took the average cumulative ground truth reward over 50 policy rollouts every 5000 RL training steps. These evaluation statistics were averaged over 5 seeds to account for randomness when learning a policy.

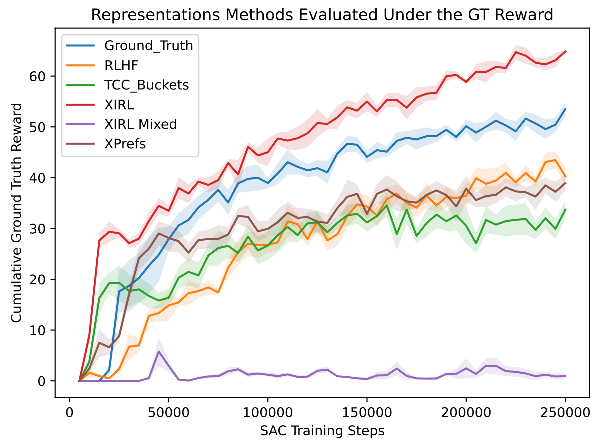


Figure 1: **Policy evaluations.** Shading denotes standard error bars. Compared with the Ground Truth reward and XIRL (optimal demonstrations) oracles, we find that XIRL using mixed quality data suffers a significant performance drop, while RLHF, XPrefs, and TCC-Buckets perform much better.

Figure 1 summarizes the results. From this graph we can infer a few key findings of this experiment. Firstly, XIRL when trained on only successful trajectories, was noticeably better than any other algorithms performance, consistently pushing all three blocks in quickly and efficiently. Interestingly, XIRL actually outperforms the ground truth reward function by learning a good representation that successfully shapes the reward function, enabling efficient RL. On the other hand, XIRL Mixed, which was the same as XIRL, but trained on an MQME dataset appears to completely collapse, unable to get a single block close to the goal zone. We hypothesize that this is because the mixed-quality data violates many of the strong assumptions that XIRL is founded on, in particular, the use of TCC to learn the latent video embedding.

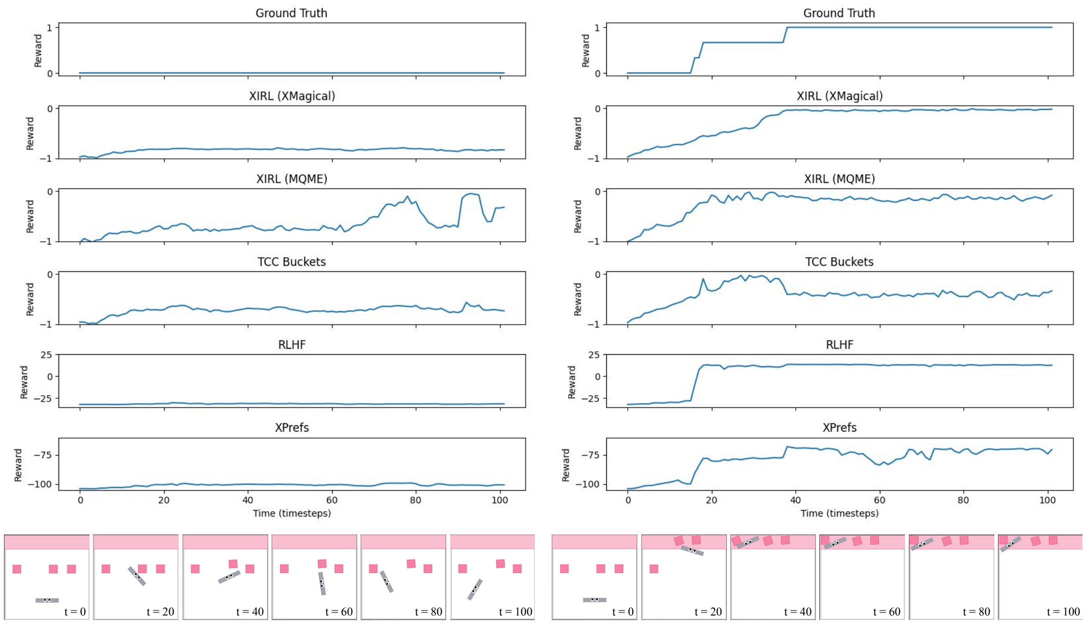
Figure 1 also shows that XPrefs, RLHF and Bucket-TCC are all quite similar quantitatively with RLHF appearing to perform slightly better and TCC-Buckets appearing to perform slightly worse. We conclude that these methods successfully leverage mixed-quality, mixed-embodiment data, achieving much better performance than the prior state-of-the-art XIRL, when also evaluated on this data. However, there still is a noticeable gap between the XIRL on near-optimal data and the MQME methods.

It is also important to note that quantitatively and qualitatively, the human effort differs across these methods. The human effort required for XIRL on X-MAGICAL for example, assumes perfect or optimal demonstrations which can be burdensome to obtain. RLHF and XPrefs on the other hand, are trained on 5000 labeled trajectories, this can prove to be a large number of times to query a human. The number of times a human would be queried for TCC-Bucket however, would be far fewer, equal to the total size of the dataset : approximately 600 categorizations across the 3 embodiments. XIRL on MQME and RL using the ground truth reward, should have little to negligible burden in comparison when it comes to the effort on the human in the loop.

6.4 Qualitative Analysis of Learned Representations and Rewards

We next performed a qualitative analysis of the learned representations and reward functions for the different methods. Figure 2 depicts the learned reward plotted against the timestep over the course of both a failed trajectory and a successful trajectory. This figure helps us visualize what reward signals and representations different approaches are learning.

Observing the shape of the learned reward for the unsuccessful trajectory, we can see that every learned reward except for XIRL trained on MQME has learned to associate actions that do not result in blocks being pushed with a low and near constant reward. The shaped reward for the successful demonstrations proves to be more informative with what these models are learning. XIRL trained on X-MAGICAL has the most shaped and dense reward, giving positive signals to the agent consistently throughout the task, even when it is moving away from the goal-zone to head back and gather more blocks. Interestingly, even XIRL on MQME has a surprisingly dense and informative signal, however this does not translate to final policy performance as it had the lowest average cumulative reward across all models. This can be explained by the fact that XIRL (MQME) likely assigns all end



(a) Learned Reward (Failed Trajectory) (b) Learned Reward (Successful Trajectory)

Figure 2: **Qualitative Analysis of Learned Rewards and Representations**. We compare the true reward with the predicted rewards across a failed **(a)** and successful **(b)** trajectory.

states as having high reward since the TCC alignment across mixed demonstrations will push the end frames of the videos are aligned to be considered successful. We can see evidence of this in the curve for XIRL (MQME) for the failed trajectory, where the predicted reward still spikes upward.

RLHF and XPrefs appear to learn a similarly shaped reward to that of the ground truth reward (a step function representing the fraction of total available blocks that are currently in the goal zone). Both methods give clear spikes in reward signals when blocks are pushed in, but fail to provide information in between these key states. On closer inspection of the policies learned by these rewards, we found they had learned a greedy strategy to get as many blocks as possible into the goal zone with a single push. The learned models do not appear to associate a strong reward signal for actions that occur after getting 2 blocks in.

7 CONCLUSION

In this paper we introduce the novel problem setting of cross-embodiment learning from mixed-quality data. Collecting near-optimal demonstrations in complex environments is challenging and human demonstrations are often noisy or suboptimal. We propose and evaluate XPrefs, RLHF and Bucket-TCC as three potential algorithms to help address these shortcomings. Our empirical results demonstrate that, XIRL (Zakka et al., 2022), the prior state-of-the-art approach to cross-embodiment IRL suffers a large degradation in performance when not all demonstrations are near-optimal. By contrast, XPrefs, RLHF and Bucket-TCC all showcase the ability to leverage mixed-quality, mixed-embodiment data to learn a reward function that is embodiment independent and enables the ability to generalize this reward to out-of-distribution embodiments. There is clear and broad scope for future work. An exciting area of future work is to explore a combination of some of our proposed methods. For example, Xprefs and Bucket-TCC appear to have somewhat complimentary reward shapes, could a linear combination of the two lead to a more informative reward? Similarly, we could leverage our goal embedding calculation and training as a finetuning step after a run of a TCC algorithm. Finally, using active preference learning could enable more label-efficient algorithms (Biyik & Sadigh, 2018; Wilde et al., 2020; Shin et al., 2023).

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 1, 2004.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Saurabh Arora and Prashant Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pp. 519–528. PMLR, 2018.

Andreea Bobu, Yi Liu, Rohin Shah, Daniel S Brown, and Anca D Dragan. Sirl: Similarity-based implicit representation learning. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 565–574, 2023.

Serena Booth, W Bradley Knox, Julie Shah, Scott Niekum, Peter Stone, and Alessandro Allievi. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5920–5929, 2023.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International conference on machine learning*, pp. 783–792. PMLR, 2019.

Daniel S Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In *Conference on robot learning*, pp. 330–359. PMLR, 2020.

Lawrence Chan, Andrew Critch, and Anca Dragan. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*, 2021.

Sungjoon Choi, Kyungjae Lee, and Songhwai Oh. Robust learning from demonstrations with mixed qualities using leveraged gaussian processes. *IEEE Transactions on Robotics*, 2019.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Caleb Chuck, Michael Laskey, Sanjay Krishnan, Ruta Joshi, Roy Fox, and Ken Goldberg. Statistical data cleaning for deep learning of automation tasks from demonstrations. In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, pp. 1142–1149. IEEE, 2017.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1801–1810, 2019.

Arnaud Fickinger, Samuel Cohen, Stuart Russell, and Brandon Amos. Cross-domain imitation learning via optimal transport. *arXiv preprint arXiv:2110.03684*, 2021.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Yang Gao, Ji Lin, Fisher Yu, Sergey Levine, Trevor Darrell, et al. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*, 2018.

Gaurav R Ghosal, Matthew Zurek, Daniel S Brown, and Anca D Dragan. The effect of modeling human rationality level on learning rewards from multiple feedback types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5983–5992, 2023.

Wonjoon Goo and Scott Niekum. One-shot learning of multi-step tasks from observation via activity localization in auxiliary video. In *2019 international conference on robotics and automation (ICRA)*, pp. 7755–7761. IEEE, 2019.

Daniel H Grollman and Aude Billard. Donut as i do: Learning from failed demonstrations. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 3804–3809. IEEE, 2011.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE symposium on security and privacy (SP)*, pp. 19–35. IEEE, 2018.

Rahul Kidambi, Jonathan Chang, and Wen Sun. Mobile: Model-based imitation learning from observation alone. *Advances in Neural Information Processing Systems*, 34:28598–28611, 2021.

Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

Yi Liu, Gaurav Datta, Ellen Novoseller, and Daniel S Brown. Efficient preference-based reinforcement learning using learned dynamics models. *International Conference on Robotics and Automation (ICRA)*, 2023.

YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1118–1125. IEEE, 2018.

R Duncan Luce. *Individual choice behavior: A theoretical analysis.* Courier Corporation, 2005.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning*, pp. 1678–1690. PMLR, 2022.

Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning multimodal rewards from rankings. In *Conference on Robot Learning*, pp. 342–352. PMLR, 2022.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pp. 278–287, 1999.

Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pp. 2, 2000.

Haoyi Niu, Jianming Hu, Guyue Zhou, and Xianyuan Zhan. A comprehensive survey of cross-domain policy transfer for embodied agents. *arXiv preprint arXiv:2402.04580*, 2024.

Jo Ann Oravec. Rage against robots: Emotional and motivational dimensions of anti-robot attacks, robot sabotage, and robot bullying. *Technological Forecasting and Social Change*, 189:122249, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

Jongjin Park, Younggyo Seo, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. Surf: Semi-supervised reward learning with data augmentation for feedback-efficient preference-based reinforcement learning. In *International Conference on Learning Representations*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Tim Salimans and Richard Chen. Learning montezuma's revenge from a single demonstration. *arXiv preprint arXiv:1812.03381*, 2018.

Roger N Shepard. Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345, 1957.

Kyriacos Shiarlis, Joao Messias, and Shimon Whiteson. Inverse reinforcement learning from failure. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 2016.

Daniel Shin, Anca D Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

Thomas Tian, Chenfeng Xu, Masayoshi Tomizuka, Jitendra Malik, and Andrea Bajcsy. What matters to you? towards visual representation alignment for robot learning. In *The International Conference on Learning Representations (ICLR)*, 2024.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S Brown. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2018a.

Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018b.

Faraz Torabi, Garrett Warnell, and Peter Stone. Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*, 2019.

Albert Wilcox, Ashwin Balakrishna, Jules Dedieu, Wyame Benslimane, Daniel Brown, and Ken Goldberg. Monte carlo augmented actor-critic for sparse reward deep reinforcement learning from suboptimal demonstrations. *Advances in Neural Information Processing Systems*, 35:2254–2267, 2022.

Nils Wilde, Dana Kulić, and Stephen L Smith. Active preference learning using maximum regret. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10952–10959. IEEE, 2020.

Nils Wilde, Erdem Bıyık, Dorsa Sadigh, and Stephen L Smith. Learning reward functions from scale feedback. *arXiv preprint arXiv:2110.00284*, 2021.

Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017.

Marty J Wolf, K Miller, and Frances S Grodzinsky. Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47 (3):54–64, 2017.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pp. 3536–3555. PMLR, 2023.

Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *Advances in neural information processing systems*, 32, 2019.

Kevin Zakka, Andy Zeng, Pete Florence, Jonathan Tompson, Jeannette Bohg, and Debidatta Dwibedi. Xirl: Cross-embodiment inverse reinforcement learning. In *Conference on Robot Learning*, pp. 537–546. PMLR, 2022.

Jiangchuan Zheng, Siyuan Liu, and Lionel M Ni. Robust bayesian inverse reinforcement learning with sparse behavior noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2198–2205, 2014.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

A RLHF Details

In this approach, the demonstration dataset is augmented by a set of pairwise preference labels over the data. For a pair of demonstrations, (v_i, v_j), the notation $v_i \prec v_j$ indicates a preference of demonstration j over demonstration i. The final form of the data is the triple (v_i, v_j, μ), where $\mu \in \{(1,0),(0,1)\}$ represents the human's preference label. For our problem, it is worth noting that the preferences include mixed-embodiment preferences, i.e., v_i may be demonstrated by one embodiment and v_j may be from a different embodiment.

Using these labels, we follow prior work by employing a deep neural network, namely a reward predictor, $\hat{r}$, that maps video frames into a single real-valued reward. Following the Bradley-Terry (Bradley & Terry, 1952) and Luce-Shepperd (Luce, 2005; Shepard, 1957) model,

$$P(v_1 \prec v_2) = \frac{exp \sum_{s \in v_2} \hat{r}(s)}{exp \sum_{s \in v_1} \hat{r}(s) + exp \sum_{s \in v_2} \hat{r}(s)}$$

where P is the softmax probability that $v_1 \prec v_2$ based on $\hat{r}$. The learned reward function, $\hat{r}$, is then optimized using a Cross Entropy Loss between the predicted value of P and the preference labels:

$$\mathcal{L}(\hat{r}) = - \sum_{(v_1, v_2, \mu) \in D} \mu(1)\ logP(v_2 \prec v_1) + \mu(2)\ logP(v_1 \prec v_2)\ .$$

B Bucket-TCC Details

Pseudo-code for Bucket-TCC is provided in Algorithm 1

Algorithm 1 Bucket TCC

Require: $D = [\tau_0, \tau_1, \cdots, \tau_n]$, bucketSize $\xi \geq 2$
Ensure: $\tau_i \prec \tau_j \;\forall\; i < j$ ▷ Ordinal Input Rankings
  $\phi \leftarrow$ TCCModel()
  $B \leftarrow$ PartitionData(D, ξ) ▷ Create $\lceil \frac{|D|}{\xi} \rceil$ buckets
  for $b \in B$ **do**
    for all pairwise $(\tau_1, \tau_2) \in b$ **do**
      $\phi \leftarrow \phi.align(\tau_i, \tau_j)$ ▷ TCC Loss aligns τ_i, τ_j
    end for
  end for

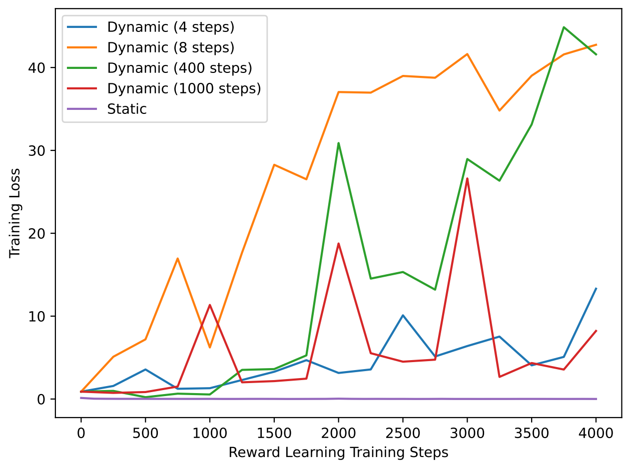


Figure 3: **Static vs. Dynamic Goal Embeddings for XPrefs.** We find that using dynamic goal embeddings that were periodically updated every X steps during training leads to training instabilities, hindering good representation learning. By contrast, using a static goal embedding leads to a convergent training loss.

C Static vs. Dynamic Goal Embeddings for XPrefs

One of the design choices mentioned in the section describing Xprefs, is whether to use a static or dynamic goal embedding. At first glance, a dynamic embedding seemed to be the best way to ensure the guarantee of a global minima so we performed a preliminary experiment with varying frequencies of goal embedding updates in the training process. Figure 3 showcases the training losses over training steps we observed for the various frequencies of updating the goal embedding g described in the XIRL preliminaries. We experimented with frequencies of 4, 8, 400 and 1000 and compared the results with using static goal embedding frozen before training. Our results provide evidence that periodically updating the goal embedding causes instabilities since it induces a moving target during learning.

In Figure 3, the loss curve when training XPrefs and updating the goal embedding every 1000 steps is particularly interesting as there is a clear spike in loss every time the goal embedding is updating, leading to an intuition that the updates lend to an instability and make it more difficult to settle in a local minimum. Our results provide evidence that periodically updating the goal embedding causes instabilities since it induces a moving target during learning. We settled on using a static embedding model which resulted in a much more stable and meaningful learned reward, as shown by the clear convergence of the loss in Figure 3 when using a static goal embedding. We theorize that fixing an arbitrary point in the embedding space as our goal state is appropriate and in fact beneficial to learning an embodiment-independent state representation as it is akin to fixing an origin for the space and fitting the state distribution around it in a way that is locally optimal.